APPENDIX A
METHODS FOR IDENTIFYING DATA OUTLIERS

To improve data quality, we flagged values for rate of use which are so large they are probably errors. Errors occur, for example, when those reporting pesticide use shift decimal points during data entry. We used three different criteria to identify outliers by comparing each use rate with an estimate of the maximum rate for that type of use.

Rate of use is not one of the fields in the PUR table. Rates are calculated by dividing the pounds of pesticide used by the acres or unit treated. Thus, an error in rate of use could occur through an error in recording either pounds or unit treated.

Only extremely large rates are flagged, not extremely small ones, because only large values will have a major influence on statistics involving pounds of pesticide use. What value to use for the maximum rate in each criterion is somewhat arbitrary; the value determines how conservative one wants to be. We chose maximum rates to be close to what were considered obvious outliers by a group of scientists in a survey described below in the description of the neural network criteria.

There are many possible methods for determining if a value is an outlier. If we knew the maximum label rates for particular uses, then rates in the PUR could be compared to these maximum rates, but unfortunately this information is not available in the PUR or in the Pesticide Label Database. The other methods to identify outliers involve looking at the distribution of the actual use rates. If the values are normally distributed, then one can identify outliers using a number of statistical procedures. If the values have an unknown or nonstandard distribution, then there exist no standard statistical procedures for identifying outliers. Nevertheless, people can look at a distribution and usually say with different degrees of confidence whether some value is an outlier. This suggests there should be some kind of procedure that can be developed to make similar judgments.

For most of the pesticide use data, distributions of rates are not even close to normal. They may have several different peaks (multi-modal). They can have either very broad or very narrow distributions. None of the standard statistical measures of outliers are very useful for these data. The best single method is the one based on neural networks. However, each different criterion will catch different outlier values so it is usually best to use all three criteria. It should be noted that these criteria are not perfect. They are conservative, meaning a value must very extreme to be flagged and so they will miss some errors. On the other hand, they may occasionally flag an extreme value that is actually correct. Because the criteria are conservative these later kinds of errors are minimized.

Criterion 1: Pounds per acre of active ingredient is larger than 200 (for non-fumigants), or 1000 (for fumigants).

Records were flagged in the PUR by criterion 1 if the pounds per acre of a non-fumigant active ingredient were greater than 200 or if the pounds per acre of a fumigant active

ingredient were greater than 1000 (column ai_a_1000_200 in the outlier table). These limit values were chosen based on what is known about typical rates of use for most pesticides.

Note that this criterion uses the pounds of active ingredient. Also, this criterion only applies to records where the unit treated is acres. The other criteria use pounds of pesticide product and apply to any unit treated, such as square feet or cubic feet.

Criterion 2: Pounds per unit treated of a product is larger than 50 times the median.

Records were flagged by criterion 2 if the pounds of pesticide product per unit treated were greater than 50 times the median value of all rates with similar types of use (column prd_u_50m in the outlier table). The median, like the mean (average), is a measure of the location of a set of values and is defined as the value in the set that has an equal number of values above and below it. It was used rather than the mean because it is not as likely to be affected by a few extreme outliers. The median was calculated from the set of all use rates of the same pesticide product and uses as that of each record being examined. By the same uses, we mean the uses of a product on the same crop or site, same unit treated, and same record type. A record type is basically either an agricultural or non-agricultural use.

Criterion 3: Pounds per unit of product is larger than a value generated using a neural network.

Records were flagged by criterion 3 if the pounds of a pesticide product per unit treated were greater than a limit value calculated using a neural network procedure (column nn4 in the outlier table).

A neural network is a special kind of function that calculates a set of output values from a set of input values. This function has a large number of parameters that must be determined so that the function will give the correct outputs for every possible set of inputs. The values for these parameters are found by a training procedure that involves presenting to the neural network program data consisting of many sets of input and corresponding output values. The program then adjusts the parameters in the neural network function until it produces the correct output values for each input set. Once the neural network has been successfully trained, it can then be used to produce appropriate output values for any input data set provided to it.

The data used to train the neural network used in the PUR outlier program were generated from frequency distributions of the pounds of pesticide product per unit treated for a selected set of pesticides and sites. Groups of pesticides and sites were chosen that included a wide range of types of distributions, including many unusual distributions. Two hundred frequency distributions were plotted and then these plots were examined independently by 12 scientists in DPR who marked rates on each plot they thought were outliers.

The results of this survey were summarized by finding an outlier maximum rate for each distribution. The maximum rate was set at a value where all 12 scientists thought higher rates were obvious outliers. These maximum rates were used as the output values for training the neural network. The input values were a set of statistical measures that described the frequency distributions. These sets of input and output values were used to train the neural network. After the neural network was successfully trained, it was used to find the outlier maximum rate for all sets of pesticide use types in the PUR.

For a more detailed explanation of the procedures used to identify outliers, see the report "A Computer Program to Identify Outliers in the Pesticide Use Report Database", L. Wilhoit, April 1998, DPR report PM 98-01.